Cahier des charges

Amélioration d'un programme de cartographie par séquençage: création d'une interface d'entrée/sortie.

Auteurs :
Hermes Paraqindes
Juliette Geoffray
Eric Cumunel

Encadrants :
Fabrice Besnard
Laurent Gueguen

12 Février 2018

Contents

1 Présentation du projet **2**
1.1 Contexte . . . 2
1.2 Objectifs . . . 2
1.3 Description de l'existant . . . 3
1.4 Critères d'acceptabilité du produit . . . 4

2 Expression des besoins **5**
2.1 Besoins fonctionnels . . . 5
2.1.1 Fonctionnalité du logiciel . . . 5
2.1.2 Fonctionnalité du code . . . 5
2.2 Besoins non fonctionnels . . . 5

3 Contraintes **6**
3.1 Délais . . . 6
3.2 Autres contraintes . . . 6

4 Déroulement du projet **6**
4.1 Planification . . . 6
4.2 Plan d'assurance qualité . . . 6
4.3 Responsabilités . . . 7
4.3.1 Maîtrise d'ouvrage . . . 7
4.3.2 Maîtrise d'œuvre . . . 7

5 Bibliographie **7**

1 Présentation du projet

1.1 Contexte

La mutagenèse est un processus par lequel l'information génétique d'un organisme, et donc celle de son ADN est modifiée, entrainant une mutation.
L'apparition des mutants dans une population est un phénomène rare et possède donc une très faible probabilité de se produire. Afin d'augmenter cette probabilité, les organismes peuvent être traités par des agents mutagènes qui vont induire des mutations (remplacement, modification ou endommagement) sur l'organisme de manière aléatoire. Ces agents mutagènes peuvent être de nature chimique, comme l'EMS (Méthanesulfonate d'éthyle),de nature physique, comme la lumière ultraviolette et les radiations ionisantes, ou de nature bactérienne pathogénique contenant des plasmides (ex: T DNA) capables de d'intégrer au génome de l'hôte une séquence ADN (capable de s'exprimer ou non).
Provoquer des mutations dans un organisme d'intérêt permet de localiser les gènes d'intérêts, de les cartographier et d'en déduire des informations sur le rôle des gènes. L'identification des mutations responsables du phénotype du mutant constitue le principe de base de la génétique mendélienne. La méthode utilisée pour identifier la mutation recherchée est celle du clonage positionnel.
Cette méthode est utilisée dans les cas où l'on ne connaît ni la séquence, ni la fonction du gène mais dont la mutation est supposée être à l'origine d'un caractère phénotypique visible. Un croisement avec une lignée génétiquement différente n'ayant pas le phénotype mutant est réalisé. Suite au croisement, il est nécessaire de génotyper un grand nombre d'individu et cela est une tâche longue et fastidieuse. La révolution dans les nouvelles technologies de séquençage et d'assemblage du génome a cependant facilité le processus. Aujourd'hui nous pouvons séquencer plusieurs mutants simultanément et détecter des variations génomiques sur tout le génome en une seule fois.
Toutefois, ce processus est une analyse bio-informatique qui requiert de nombreuses étapes ainsi que l'utilisation de plusieurs programmes. La plupart des programmes utilisés sont adaptés à un organisme modèle ou à un « design » génétique, et dépendent de serveurs distants. Afin de faciliter l'étape de cartographie et l'identification des mutations, un pipeline appelé *Andalusian_ Mapping* a été développé. Ce pipeline permet de travailler avec différentes espèces et souches de cartographie.

1.2 Objectifs

L'objectif général de ce projet est de rendre *Andalusian_ Mapping* plus accessible à la communauté scientifique, notamment pour les biologistes non-informaticiens, à la fois au niveau de la mise en place des outils du pipeline que de l'affichage des résultats.
Pour ce faire, plusieurs pistes peuvent se révéler intéressantes :

- Au niveau de la mise en place : incorporer le programme dans un docker contenant les dépendances logicielles nécessaires.
- Au niveau de l'importation des données de départ : intégrer une interface graphique permettant d'ajouter les fichiers plus aisément.
- Au niveau de l'exploitation des résultats finaux : créer une interface graphique et/ou une page HTML pour visualiser les différents résultats (sous la forme de graphes et de tableau).
- Améliorer l'organisation du pipeline (optionnel).

1.3 Description de l'existant

Récemment, Cloudmap, un pipeline automatique de mapping-by-sequencing et d'identification de mutants a été développé et intégré à Galaxy, qui est une interface utilisateur simple et intuitive. Cependant, ce pipeline ne permet pas à ce jour de travailler avec des génomes de référence *exotiques*, étant donné que seuls les gènomes modèles comme *Caenorhabditis elegans* ou *Arabidopsis thaliana* sont disponibles sur Galaxy. *Andalusian Mapping* a donc été développé dans le but de fournir un outil similaire afin de travailler avec une gamme plus vaste d'organismes.
Andalusian Mapping est un pipeline permettant d'effectuer une cartographie par séquençage afin d'identifier les régions les plus susceptibles d'être responsables d'une mutation. Ce pipeline est constitué d'un script bash permettant l'utilisation en une seule étape de plusieurs outils bioinformatiques.
Il tourne sous environnement Linux et Mac. La mise en place de différents logiciels est cependant fastidieuse et nécessaire au fonctionnement de ce pipeline.
Logiciels requis :

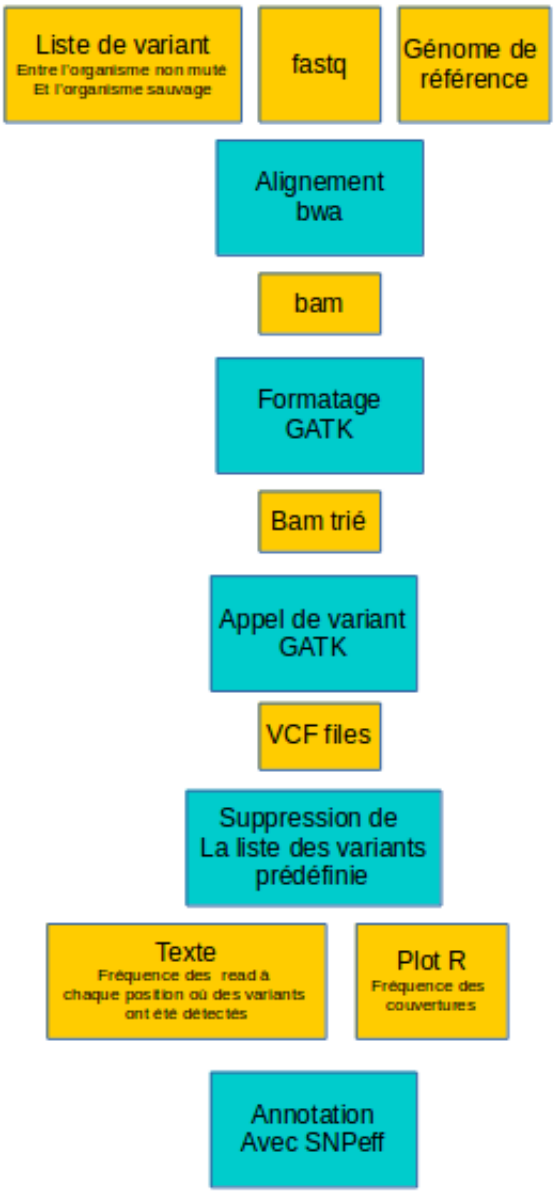
- bwa version 0.7.5a-r405 ou supérieur
- samtools version 0.1.18 ou supérieur
- Picard Version 1.110 ou supérieur
- GATK Version 3.7 ou supérieur
- R Version 3.3 ou supérieur, avec le package ggplot2.
- snpEff Version 4.1g ou supérieur

L'utilisation de ce pipeline se base sur un fichier de configuration à remplir préalablement. Celui-ci contient les chemins vers les exécutables des logiciels utilisés, les fichiers *fastq*, le génome de référence, une liste de variants prédéfinie ainsi qu'un répertoire dans lequel seront envoyés les résultats intermédiaires et finaux.

Description du pipeline dans son ensemble :

1. Mapping et sorting des lectures contre le gènome de référence avec BWA
2. Ajout des informations de groupe de lectures
3. Merge des duplicats de lecture détectés
4. Indexation du fichier BAM
5. Génération de statistiques d'alignement (fréquence, couverture)
6. Réalignement des lectures en prenant en compte les informations récoltées jusqu'ici
7. Base Quality Score Recalibration avec génération de graphiques et de statistique concernant la qualité des alignements
8. Variant Calling

9. Réalisation de la cartographie des mutations à l'aide des SNP connus pour chaque scaffold
10. Détection des meilleurs locus candidats à l'aide des annotations.
11. Génération des graphiques à l'aide d'un script R et d'un fichier contenant les locus les plus probables d'être responsables de la mutation



1.4 Critères d'acceptabilité du produit

Afin de répondre à la problématique, le pipeline sera remanié et une interface graphique sera développée. Elle permettra de contourner l'étape fastidieuse du remplissage du formulaire, renseigner l'utilisateur du bon déroulement de l'exécution du pipeline et enfin, faciliter la visualisation des résultats.

2 Expression des besoins

2.1 Besoins fonctionnels

2.1.1 Fonctionnalité du logiciel

La fonctionnalité du pipeline reste inchangée : il effectuera une cartographie automatisée des mutations, mais il sera désormais accompagné d'une nouvelle interface graphique plus intuitive.
De plus, l'installation des dépendances devra être facilitée grâce à l'implémentation d'une plateforme de type Docker.
Enfin, des conditions et des vérifications d'existence de fichiers seront ajoutées au script afin de rendre l'exécution du pipeline plus robuste.

2.1.2 Fonctionnalité du code

Le code apportera les fonctionnalités suivantes :

- Résolution des problèmes de dépendances logicielles grâce à l'implémentation d'une plateforme de type Docker
- Remplissage des champs du fichier de configuration du pipeline via une interface graphique :
  - Génome de référence
  - Fastq en pairend
  - Liste d'allèles connus entre l'individu sauvage et l'individu qui sera étudié, avant mutation (pour pouvoir les supprimer lors de l'appel de variant entre l'organisme sauvage et l'organisme muté.)
  - Données d'information sur les reads (comme la technique de séquençage utilisée, le SNP utilisé pour repérer la région mutée...)
  - Le nom du dossier de sortie et de l'extension ajoutée au fichier.
- Remplissage automatique de certains champs via des pipelines développés par le maître d'ouvrage (comme la génération de la liste des variant entre l'organisme sauvage et l'organisme étudié avant mutation)
- Visualisation claire des données et des graphiques d'intérêt.
- Un fichier de sortie exportable (probablement sous le format d'une page html), pour faciliter le partage de ces résultats.

2.2 Besoins non fonctionnels

L'interface développée sera réalisée sous Linux, il comprendra des parties en Shell, en Python et nous utiliserons également *pyQt*, une librairie graphique python.

3 Contraintes

3.1 Délais

Le projet s'organise sur une période de 4 semaines. Il débutera le lundi 5 février et prendra fin le vendredi 2 mars. Le projet pourra être poursuivi durant la période de cours allant du 5 au 29 mars.
Un cahier des charges définitif sera délivré le lundi 12 février.
Le délivrable final sera rendu le 29 mars.

3.2 Autres contraintes

4 Déroulement du projet

4.1 Planification

Listes des taches :	Distribution des rôles			5 au 9	12 au 16	19 au 23	26 au 2	5 au 29
	Juliette	Hermes	Eric	Février – Mars				
Rendez-vous avec le référent								
Rédaction du cahier des charges								
- Introduction								
- Existant								
- Autre partie								
Installation et vérification de la bonne marche du pipeline								
Réécriture du Pipeline								
- Ajout de condition a chaque création de fichier								
- Vérification des fichiers créer								
Mise en place d'un docker								
- Lecture de la documentation								
- Mise en place du docker								
Interface graphique								
Développement de la partie formulaire								
- Formulaire simple								
- Formulaire interactif								
Développement de la partie recharge d'un formulaire								
Développement de la partie exécution du pipeline								
Développement de la partie visualisation des résultats								
Développement de la partie recharge d'ancien résultats								
Beta testing								
Documentation								

4.2 Plan d'assurance qualité

Afin de nous assurer de la pertinence et de la bonne qualité du programme développé, nous ferons en sorte d'obtenir un logiciel utilisable au plus tôt. Puis, en tenant compte des retours du maître d'ouvrage, de rajouter et d'optimiser les différentes fonctionnalités recherchées.
Le programme final devra être simple d'utilisation, robuste et devra informer l'utilisateur en cas d'erreurs.
Enfin, celui-ci sera testé à la fois par le maître d'ouvrage (Dr Fabrice Besnard) et par des utilisateurs non-initiés en informatique.

4.3 Responsabilités

4.3.1 Maîtrise d'ouvrage

Ce projet nous a été proposé par Fabrice Besnard, biologiste rattaché à l'École Normale Supérieur de Lyon (ENS), travaillant au Laboratoire de Reproduction et Développement des Plantes (RDP).

4.3.2 Maîtrise d'œuvre

Pour réaliser ce projet, nous serons trois étudiants en Master 1 de Bio-informatique à Lyon 1 : Juliette Geoffray, Hermes Paraqindes et Eric Cumunel.

5 Bibliographie

1. Schneeberger, K. Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat. Rev. Genet. 15, 662–676 (2014).

2. Galaxy | Published Page | CloudMap. Available at: https://usegalaxy.org/u/gm2123/p/cloudmap. (Accessed: 5th September 2017)

3. Home of SHOREmap. Available at: http://bioinfo.mpipz.mpg.de/shoremap/index.html. (Accessed: 5th September 2017)

4. MutMap - Genomics & Breeding. Available at: http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap. (Accessed: 5th September 2017)

5. Besnard, F., Koutsovoulos, G., Dieudonné, S., Blaxter, M. & Félix, M.-A. Toward Universal Forward Genetics: Using a Draft Genome Sequence of the Nematode Oscheius tipulae To Identify Mutations Affecting Vulva Development. Genetics 206, 1747–1761 (2017).

6. Besnard, F. Andalusian_Mapping: Scripts to perform mapping-by-sequencing. (2017).